Biostatistical Review

BLA:	**STN# 103780/0** *Rebif® for the treatment of relapsing remitting forms of Multiple Sclerosis (RRMS)* Submission received September 5, 2001 Serono, Inc. Norwell, MA
Date:	December 27, 2001
Reviewer:	***Clare Gnecco, Ph.D.***
Through:	***Ghanshyam Gupta, Ph.D.*** ***Branch Chief, Therapeutics Group***
cc:	HFM-99/DCC: BLA #103780/0 HFM-576/Dr. Rask HFM-576/Dr. Walton HFM-588/Ms. Giuliani / Ms. Winestock FM-210/Dr. Ellenberg HFM-215/Chron – File: BLAREBIF.DOC

STATISTICAL REVIEW ISSUES / SUMMARY:

The sponsor's major efficacy and safety analyses were investigated and major statistical claims confirmed. Only those additional statistical analyses performed by this reviewer and analyses requested by the reviewing medical officer, Cynthia Rask, MD, are presented in this review.

SUMMARY OF STATISTICAL ISSUES:

(1) The sponsor's study center pooling strategy: Per the pre-specified strategy in the sponsor's statistical analysis plan (SAP), pooling of study centers for inclusion of center as a main effect in analyses was to have been based on geographic considerations for small centers. In fact, the pooling strategy actually used was data driven which is problematic. *NOTE: There were 56 participating centers from 9 countries. The smallest recruiting center had 3 subjects, 2 centers contributed 4 subjects, and 5 centers contributed 6 subjects each. The remaining centers contributed between 6 – 24 subjects each (CSR, Table 3, pp. 65-66).* This reviewer performed analyses of major efficacy endpoints based on strict geographic pooling of centers into 3 groups (US, Canada, and Europe) as well as un-pooled analyses (not including the center effect). In addition, descriptive analyses for individual centers were also performed for the primary and major secondary efficacy endpoints. The sponsor's positive statistical findings were found to be robust based on these analyses.

(2) The sponsor's finalized statistical analysis plan (SAP) defines the ITT (intent-to-treat) analysis population as all subjects randomized. However, the clinical study report (CSR) uses a different definition, viz., all those subjects randomized and who received at least one injection of open label treatment. This is really the all subjects treated population. The primary analysis was specified based on the true ITT group and this is the one that should be used. In fact, however, only one patient was randomized, but never treated.

(3) Regarding Poisson regression modeling for the major secondary efficacy endpoint of exacerbation count: (i) An analysis to verify the Poisson data distribution assumption was not provided and needed to be investigated and (ii) this reviewer also performed a nonparametric Wilcoxon rank sum test as well as a distribution-free permutation test to assess the robustness of Poisson modeling findings. The reported findings were found to be robust.

(4) There was one very large outlier (viz., a count of 83) and another large value (viz., a count of 42) for the number of baseline T1 lesions. These were confirmed to be the actual values and both occurred in Avonex® treated subjects. The impact of these values on statistical analytic results was investigated by performing stratified analyses of endpoints involving T1 lesion counts using the overall median count to define strata cutpoints.

(5) During the review period, the sponsor identified one problematic study site (in the US), which contributed 11 patients. Major efficacy analyses were performed excluding this site and the results were found to be robust.

BACKGROUND:

Rebif ® is currently approved in Europe for the treatment of relapsing-remitting forms of MS. This review will focus solely on one controlled study, XXXXXXXXXX, *"An open-label, randomized, multicenter, comparative, parallel group study of Rebif® 44 mcg*

administered three times per week by subcutaneous injection, compared with Avonex® 30 mcg administered once per week by intramuscular injection in the treatment of relapsing-remitting multiple sclerosis." This trial was conducted in 56 clinical centers (36 in the US, 5 in Canada and 15 in Europe). The sponsor's protocol synopsis, detailing the study design, is appended to this review. Also, refer to the clinical review for a full description of this study's design and conduct.

SUMMARY OF STUDY XXXXXXXXXX: The study was conducted from November 1999 to February 2001.

Study Objectives: The **primary objective** was to demonstrate that the proportion of subjects with R-R MS who were exacerbation-free is greater with Rebif ® 44 µg administered three times/week than with subjects treated with Avonex® 30 µg once/week for 24 weeks. The principal secondary objective was to demonstrate that the MRI-determined combined unique (CU) lesion activity is less after 24 weeks of treatment.

Study Endpoints:
The **primary efficacy endpoint** was the proportion of subjects who were exacerbation-free after 24 weeks.
Secondary efficacy endpoints were, as ordered prospectively by the sponsor, as follows: (i) mean number of CU (combined unique) T1 + T2 active MRI lesions per subject per scan during 24 weeks of treatment (ii) total exacerbation count per subject and (iii) mean number of T2 active lesions per subject per scan.
Tertiary Endpoints included: (i) mean number of T1 active lesions per subject per scan (ii) proportion of CU, T2, and T1 active scans per subject and (iii) proportion of subjects with no active CU, T2, and T1 lesions during the study period.
The usual **safety endpoints** were assessed as well as depression via the Beck Depression Inventory (BDI).
Randomization: Randomization to study treatment was carried out centrally via a centralized telephone system. Randomization was stratified by study center utilizing an initial block size of six followed by block sizes of four. The sponsor stated that this strategy was employed to prevent potential detection of treatment codes and they carried out simulation studies to characterize its operating characteristics.

Sample Size: The planned enrollment for this study was 624 subjects allocated in a 1:1 ratio to Rebif® or Avonex®. Although all enrolled subjects were to complete 48 weeks of treatment, efficacy outcomes were to be assessed after 24 weeks of treatment. A sample size of 280 evaluable subjects per treatment group was estimated to provide 95% power to detect a ? = 30% improvement in the primary endpoint in the Rebif ® vs. Avonex® groups. This calculation assumes a two-sided Chi-square test and a type I error of 0.05. Assumptions used were that the proportion of exacerbation-free subjects at 24 weeks was 65% on the Rebif ® arm vs. 50% on the Avonex® arm. These estimates were derived from data obtained in the PRISMS and OWIMS trials. This sample size also provides 99% power to detect a 46% reduction in the mean number of CU lesions per subject per scan during 24 weeks of treatment. For this calculation, a two-sided Wilcoxon rank-sum test at the 0.05 level of significance, assuming a common standard deviation of 0.95, was

employed. Further assumptions used were that the mean number of CU lesions per subject per scan during 24 weeks of treatment is 0.42 on the Rebif ® arm vs. 0.78 on the Avonex® arm. These estimates were derived from phase III Rebif ® studies in similar subject populations which utilized 44µg three times per week and 44µg once per week. Assuming a 10% attrition rate, 312 subjects per group or a total of 624 were to be randomized to treatment. In fact, 677 subjects were randomized, 339 to Rebif® and 338 to Avonex®). Only one subject, assigned to Avonex®, was not treated. All other subjects received their assigned treatment.

Analysis Populations: Two subject cohorts were to be analyzed: (i) an **ITT (intent-to-treat) group** defined as all subjects who were randomized and received at least one injection of open label treatment and (ii) an **evaluable group** which included those subjects who had no major protocol deviations and who had either completed 24 weeks of treatment or satisfied criteria specific to individual endpoints. **The ITT analysis was considered primary.** Because two study sites (#267 in France with 5 subjects and #291 in Canada with 22 subjects) had a priori chosen not to perform MRI scans, subjects from these two centers were excluded from the ITT efficacy population for analysis of MRI parameters. Thus, 650 ITT patients were analyzed for the MRI parameters.
Reviewer's Comment: The above-mentioned ITT definition is not true ITT, but rather all subjects treated. The finalized SAP pre-specified the true ITT group (i.e., all patients as randomized) as the primary analysis group. This is the one that must be used. Thus, denominators should include 677 subjects, not 676.

Interim Analysis: The pre-specified interim analysis (i.e., when half the subjects had either completed 24 weeks of treatment or withdrew before 24 weeks) was not performed. The purpose of this analysis was to have been possible stopping for futility and safety concerns. This interim analysis was deleted by Amendment 4, dated November 9, 2000.

STATISTICAL METHODOLOGY:

The Statistical Analysis Plan (SAP) was finalized on November 16, 2000. All statistical testing was two-sided, at the 0.05 level of significance. The main efficacy analysis was to occur when all enrolled subjects had either completed 24 weeks of treatment or had stopped treatment before 24 weeks. The **primary efficacy endpoint**, proportion of exacerbation-free subjects at 24 weeks, was analyzed using a logistic regression model adjusting for treatment and study center. The **main secondary endpoint**, average of the ranked mean number of CU active lesions per patient per scan during 24 weeks of treatment, was analyzed using a nonparametric ANCOVA with the baseline number of active lesions as the covariate and adjusting for treatment and study center. The reported adjusted treatment means and associated standard errors were estimated using the analogous parametric model. Treatment differences of the adjusted means with associated 95% CI's were also presented. All other MRI parameters, with the exception of proportions, were analyzed similarly. Exacerbation counts were analyzed using a

Poisson regression model with effects for treatment and center. Log (time on study) was used as the offset variable in this model.

Specifics of the Analyses per the Sponsor's Clinical Study Report (CSR): Those instances where changes were made since the finalized SAP are identified and described.

- **Study Center Pooling Strategy:** The finalized SAP states the following: *"In analyses adjusting for center, all centers will remain at independent levels of the center effect except with centers with less than 3 patients per treatment group. Centers with less than 3 patients per treatment group will be grouped by geographical region (i.e., US, Europe, or Canada) as three different pooled centers each being a separate independent level of the center effect. If there are less than 3 patients per treatment group in any of these pooled centers, then these patients will be pooled with the next smallest independent center in the same geographical region."* In addition to this pooling strategy, the CSR states that *"……for all main effects models, centers were pooled by geographic region if all patients in a center in both treatment groups had the same response for the dependent variable. If this pooling was not performed, then it would not have been possible to assess the treatment effect for such centers as there would have been no variability within these centers."* Operationally speaking, for the main effects model (including only Treatment and Center and no interaction term), there were 48 independent levels for the Center effect using the pre-specified pooling strategy. Whereas, for the full model (including the Treatment X Center interaction term), there were 35 independent levels for the Center effect using the amended strategy. The previously described amended pooling strategy was applied because the interaction effect could not be assessed for the full model due to obtaining a non-positive definite inverse Hessian covariance matrix of parameter estimates (the SAS statistical procedure does not converge if one uses 48 levels for Center). The sponsor considers the results of the model with 48 levels for Center as primary.

Reviewer's Comment: While the statistical rationale for this approach is understood, from a regulatory standpoint it is unacceptable as it is data dependent. This reviewer performed both an un-stratified (treatment only model) analysis and one stratified by geographic location (3 levels: US, Europe, Canada). In both cases, a statistically significant improvement in treatment effect, favoring Rebif®, was found. These results are presented in the **'REVIEWER's EFFICACY ANALYSES'** section of this review.

- For all logistic regression analyses, a full effects model (with treatment, center, and treatment X center terms) was performed in order to test for interaction effects. The sponsor reports that no statistically significant interactions were found.
- Values of MRI parameters at Study Day 1 were considered as the baseline values if both measurements at the screening visit and the Study Day 1 visit were available; otherwise, the baseline MRI parameters were considered missing observations since to determine baseline activity, both the screening and Study Day 1 scans were needed. Additionally, patients with missing baseline MRI data

had their data imputed for the ITT analysis using the overall median value at baseline for those patients who had pre-treatment measurements.

- For the efficacy-evaluable analysis for total exacerbation count and steroid use for exacerbation, the offset variable used in the Poisson regression model was the minimum value of the time to major protocol deviation and time on study for patients who completed 24 weeks; otherwise, the offset variable was the minimum value of the time to protocol deviation and the time on treatment.
- **Missing Data Imputation for the Primary Endpoint:** For subjects who withdrew before Week 24 without an exacerbation, the proportion exacerbation-free was estimated as follows: (i) The number of subjects in each treatment group who withdrew without an exacerbation was determined. (ii) The proportion of exacerbation-free subjects among those with known status (i.e., had either experienced an exacerbation before Week 24 or had completed 24 weeks without an exacerbation) was determined across both treatment groups. (iii) The number of subjects withdrawing without an exacerbation in each treatment group who would be considered exacerbation-free was determined as the product of these two numbers. These estimates were rounded up to the next integer if the decimal part was ≥ 0.5 and rounded down otherwise. This approach had been previously agreed to by CBER.
- **Missing Data Imputation for Post-baseline MRI Parameters:** If a subject had post-baseline scans, but had less than the complete set of 6, then the subject's value was estimated as follows: (i) The mean number of lesions/scan was computed using the number of scans the patient had. In other words, that number was used as the divisor. (ii) The proportion of active scans was estimated using the number of scans the subject had. If a subject's mean number of lesions/scan was 0, then the proportion of active scans was estimated as 0. Otherwise, the proportion of active scans for the subject was computed as the total number of active scans the subject had divided by the number of scans the subject had. (iii) If a subject had no post-baseline MRI scans (there were only 4 such subjects), estimation was as follows: (a) the mean number of lesions/scan was estimated as the median of the mean number of lesions/subject/scan across both treatment groups. (b) If the subject's estimated mean number of lesions/scan was 0, then the proportion of active scans was estimated as 0. If the subject's estimated mean number of lesions/scan was > 0, the proportion of active scans was estimated to be the median of the proportion of active scans/subject across both treatment groups using the data from all subject's with post-baseline MRI scans. This approach had been previously agreed to by CBER.
- **Sensitivity analyses** were performed for the ITT population analysis of the primary parameter and the main secondary parameter at baseline and during the study. A very conservative sensitivity analytic approach was used assigning all patients in the Rebif® group an exacerbation response and all patients in the Avonex® group a response of exacerbation-free. Applying the logistic regression analysis to these data yielded a highly statistically significant p-value of 0.0055. This same approach was applied in the stratified Cochran Mantel-Haenszel test (p=0.0074) as well as Fisher's exact test (p=0.0101). All p-values for these sensitivity analyses favored Rebif®.

- ***Reviewer's Comment*:** A small number of subjects withdrew from treatment, but not from study. There was a grand total of 11 treatment dropouts, 9/339 (2.7%) on Rebif® and 2/338 (0.6%) on Avonex®. The distributions by treatment arm for time on study (in days) and time on treatment (in days) were very similar. Thus, missing data imputation in this case did not prove to be problematic, as evidenced by the sensitivity analysis result.

Distribution of Time on Treatment (in Days)		
	Rebif®	Avonex®
Mean	165.2	166.1
Std. Deviation	20.57	18.47
Median	169.0	169.0
Minimum	1.0	1.0
Maximum	197.0	195.0
Sample Size	339	337

Distribution of Time on Study (in Days)		
	Rebif®	Avonex®
Mean	166.7	166.8
Std. Deviation	16.89	16.53
Median	169.0	169.0
Minimum	1.0	1.0
Maximum	197.0	195.0
Sample Size	339	337

REVIEWER's EFFICACY ANALYSES:

Primary Efficacy Endpoint: The pre-specified statistical analysis for the primary endpoint was logistic regression modeling adjusting for study site. Due to the data dependent pooling strategy, this reviewer investigated two logistic regression models. The first contained only the treatment effect. The resultant p-value was statistically significant with p=0.023. The second model included treatment and geographic location at three levels (US, Canada, and Europe). The treatment X geographic location interaction term was not statistically significant. The main effects model with treatment and geographic location terms yielded a statistically significant p-value of 0.024. In addition, this reviewer performed several ITT analyses, both adjusted and unadjusted, to examine the robustness of the sponsor's statistically significant result. Analyses were performed with and without the problematic center, #238.

(1) The **reviewer's first analysis** is an unadjusted contingency table analysis. The Fisher's exact test p-value for the following cross-tabulation is 0.0012.

	REBIF	AVONEX
	N = 339	N = 338
Exacerbation-free	254 (74.9%)	214 (63.3%)
Not Exacerbation-free	85 (25.1%)	124 (36.7%)
	Treatment Comparison	p = 0.0012
Odds Ratio (OR)	1.7	
95% CI	1.2, 2.4	
Relative Risk (RR)	1.5	
95% CI	1.2, 1.8	

When one excludes the problematic study site #238, the Fisher's exact p-value is 0.0004.

(2) The **reviewer's second set of ITT analyses** looked into consistency of finding by **stratifying** for age (< 38 years vs. = 38 years; cutpoint based on the median) and gender. The two-sided stratified Cochran Mantel-Haenszel (CMH) test was used.

Age < 38 years:

	REBIF	AVONEX
	N = 157	N = 180
Exacerbation-free	110 (70%)	108 (60%)
Not exacerbation-free	47 (30%)	72 (40%)

Age = 38 years:

	REBIF	AVONEX
	N = 182	N = 158
Exacerbation-free	144 (79%)	106 (67%)
Not exacerbation-free	38 (21%)	52 (33%)

The CMH p-value is 0.0017. The estimates for overall odds ratio and relative risk based on this stratified analysis are:

Odds Ratio: 1.7 95%CI: [1.2, 2.4]
Rel. Risk: 1.4 95%CI: [1.1, 1.8]

Excluding site #38 the p-value is 0.0006.

Males:

	REBIF	AVONEX
	N = 85	N = 86
Exacerbation-free	69 (81%)	56 (65%)
Not exacerbation-free	16 (19%)	30 (35%)

Females:

	REBIF	AVONEX
	N = 254	N = 252
Exacerbation-free	185 (73%)	158 (63%)
Not exacerbation-free	69 (27%)	94 (37%)

The CMH p-value is 0.0011. The estimates for overall odds ratio and relative risk based on this stratified analysis are:

Odds Ratio: 1.7 95%CI: [1.2, 2.4]
Re l. Risk: 1.5 95%CI: [1.2, 1.8]

Excluding site #38 the p-value is 0.0003.

(3) Reviewer's Assessment of Potential Study Trends: To assess potential changes over time in study conduct, this reviewer performed a stratified analysis of the primary endpoint by the first and second halves of the study. Patients were sorted by date of first dose for this analysis. Analytic results indicate:

First Half of Study:

	REBIF	AVONEX
	N = 172	N = 166
Exacerbation-free	132 (77%)	103 (62%)
Not exacerbation-free	40 (23%)	63 (38%)

Second Half of Study:

	REBIF	AVONEX
	N = 167	N = 172
Exacerbation-free	122 (73%)	111 (65%)
Not exacerbation free	45 (27%)	61 (35%)

The CMH p-value is 0.0011. Excluding site #38 the p-value is 0.0004.

(4) Reviewer's Geographic Stratification Analysis:

United States:

	REBIF	AVONEX
	N = 223	N = 220
Exacerbation-free	173 (78%)	144 (65%)
Not exacerbation-free	50 (22%)	76 (35%)

Canada:

	REBIF	AVONEX
	N = 35	N = 38
Exacerbation-free	24 (69%)	24 (63%)
Not exacerbation-free	11 (31%)	14 (37%)

Europe:

	REBIF	AVONEX
	N = 81	N = 80
Exacerbation-free	57 (70%)	46 (58%)
Not exacerbation-free	24 (30%)	34 (42%)

The CMH p-value is 0.0011. Excluding site #38 the p-value is 0.0004.

Impact of Baseline MRI Lesion Status on the Primary Efficacy Endpoint: This reviewer performed several ancillary stratified analyses to assess the robustness of the sponsor's findings for the primary endpoint applying a categorical adjustment for baseline CU, T1, and T2 lesion counts. In these analyses, strata for baseline lesion counts of each type were constructed using the particular overall median baseline lesion count (i.e., for both treatment groups combined). One of the reasons this approach was taken was that there were two outliers (viz., baseline T1 lesion counts of 42 and 83) which occurred in two Avonex®-treated subjects. Analyses were also performed excluding the problematic center, #238.

Baseline CU Lesion Count £ 1:

	REBIF	AVONEX
	N = 219	N = 217
Exacerbation-free	176 (80%)	140 (65%)
Not exacerbation-free	43 (20%)	77 (35%)

Baseline CU Lesion Count > 1:

	REBIF	AVONEX
	N = 106	N = 108
Exacerbation-free	68 (64%)	66 (61%)
Not exacerbation-free	38 (36%)	42 (39%)

The two-sided Cochran Mantel-Haenszel test yields a p-value of 0.0013. Excluding site #238 yields a p-value of 0.0004.

Using a cutpoint of zero yields another statistically significant result:

Baseline CU Lesion Count = 0:

	REBIF	AVONEX
	N = 146	N = 147
Exacerbation-free	116	96
Not exacerbation-free	30	51

Baseline CU Lesion Count > 0:

	REBIF	AVONEX
	N = 179	N = 178
Exacerbation-free	128	110
Not exacerbation-free	51	68

The two-sided Cochran Mantel-Haenszel test yields a p-value of 0.0012. Excluding site #238 yields the same statistically significant p-value.

Baseline T1 Lesion Count £ 0:

	REBIF	AVONEX
	N = 186	N = 178
Exacerbation-free	149 (80%)	114 (64%)
Not exacerbation-free	37 (20%)	64 (36%)

Baseline T1 Lesion Count > 0:

	REBIF	AVONEX
	N = 139	N = 147
Exacerbation-free	95 (68%)	92 (63%)
Not exacerbation-free	44 (32%)	55 (37%)

The two-sided Cochran Mantel-Haenszel test yields a p-value of 0.0014. Excluding site #238 yields a p-value of 0.0005.

Baseline T2 Lesion Count £ 0:

	REBIF	AVONEX
	N = 201	N = 205
Exacerbation-free	157 (78%)	136 (66%)
Not exacerbation-free	44 (22%)	69 (34%)

Baseline T2 Lesion Count > 0:

	REBIF	AVONEX
	N = 124	N = 120
Exacerbation-free	87 (70%)	70 (58%)
Not exacerbation-free	37 (30%)	50 (42%)

The two-sided Cochran Mantel-Haenszel test yields a p-value of 0.0011. Excluding site #238 yields a p-value of 0.0003.

Reviewer's Comment: All of the above analyses indicate that the sponsor's claim for a statistically significant improvement in the primary efficacy endpoint favoring Rebif® is indeed robust.

Secondary Efficacy Endpoints:

(1) **Exacerbation Count Endpoint:** In order to assess the robustness of Poisson regression modeling findings, this reviewer performed a nonparametric Wilcoxon rank sum test as well as a permutation test (sampling with replacement). These yielded statistically significant findings of $p = 0.003$ and $p < 0.001$, respectively.
(2) **Mean Number of CU and T2 Active Lesions Per Scan Per Patient:** The sponsor's analytic results were confirmed.

Reviewer's Tertiary Analyses:

(1) **Change in EDSS Score from Baseline to 24 Weeks:** An exploratory analysis was performed on change in EDSS score from baseline to 24 weeks. The Wilcoxon rank sum test yielded a statistically significant p-value of 0.041, favoring Rebif®. This present analysis group is contaminated by subjects who had exacerbations at six months. XXXXXXXXXX.

Subjects Who Experienced a Clinical Exacerbation Within 3 Months: There were 120 subjects in this category, 53 in the Rebif® group (15.6%) and 67 in the Avonex® group (19.8%). The distributions of time to first exacerbation were similar for the two treatment groups. For the Rebif® group the median was 48 days with a range from 1 to 89 days; for the Avonex® group the median was 48 days with a range from 2 to 89 days.

OVERALL SUMMARY AND CONCLUSIONS:

This reviewer's analyses of the major efficacy endpoints, based on the electronic database provided, confirm the sponsor's major statistically significant findings.